Semantically Relevant Scene Detection Using Deep Learning

Dipanita Chakraborty*, Werapon Chiracharit†1 and Kosin Chamnongthai†2
Department of Electronic and Telecommunication Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi, Bangkok, Thailand
*E-mail: dipanita.chakraborty@mail.kmutt.ac.th
†1E-mail: werapon.chi@kmutt.ac.th
†2E-mail: kosin.cha@kmutt.ac.th

Abstract— **Automatic detection of semantically relevant scene will make content-based video browsing, video retrieval, and video indexing tasks faster and efficient. Categorizing similar meaningful scenes of a large video is in high demand in video processing to understand the correlation between different scenes of a video, for example, which scenes have fighting action concept, or in which scenes similar criminal has appeared. Previous researches have shown supervised methods for semantic scene or concept understanding of a video; however, their method is unable to find out semantic relevance between different scenes. In this paper, a CNN-deeper LSTM based method is proposed which consists of three steps. Firstly, input video dataset is segmented into shots; secondly, features are extracted from candidate video frames, which are considered as feature descriptors; finally, these features descriptors are analyzed and evaluated for recognizing semantically relevant scenes. Experimental results show efficiency of the proposed method of this paper.**

I. INTRODUCTION

One of the post-covid-19 effects is significant increment of digital media users, which leads to large amount of video data. These large amount of dataset makes it challenging to perform video indexing, video retrieval and content-based video browsing tasks, and it even becomes more difficult to do automatic semantic concept or scene detection which is a hot topic. Automatic semantic concept or scene detection of a video understands the concept or content of a video such as birthday party or sports or national ceremony by using low-level features of the video, thus helping users to search for a particular desired events or content-based video browsing from a large video dataset. A large video consists of many scenes, and each scene defines different concepts.

This paper aims for a new approach that is to detect all those scenes which have similar concepts i.e., semantically relevant scenes from a large video dataset, such as all the sports scenes from a movie or all the president giving speech scene from a large video of a country's national ceremony. Automatic semantic video concept or scene detection mainly has two steps.

The first step focuses on low-level feature extraction from the video. After the extraction features are stored which is called descriptors. As low-level feature extraction from the video frames, traditional research works have mainly focused on visual, textual, motion vectors and audio features. In paper [1], authors have used SURF, ORB and BRISK algorithm to extract visual features. In paper [2], visual features (color information, shape) are extracted using HSV and Hu. However, the use of low-level visual feature extraction generates a semantic gap between low level visual features and semantic concept or scene detection. Furthermore, low level visual features are sensitive to video content transformation. To overcome this issue, some papers have used supervised machine learning in order to extract semantic information from the video. In paper [4], [6], and [9] authors have used CNN, SVM, and RNN to extract multiple semantic features respectively. Textual information extraction using Term-Frequency Inverse Documents (TF-ID) algorithm has been used in paper [3]. Since, motion vectors are an important key feature of a video, thus, in paper [8], HMM (Hidden Markov Model) based framework is used to extract motion descriptors from the video. Adaptive threshold is used to extract motion descriptors in paper [5]. In paper [7], global motion estimation has been used to detection motions vectors as feature descriptors.

The final step of semantic scene detection is to match low-level feature descriptors with the high-level semantic concepts of the scenes. Since machine learning algorithms are great for feature recognition and classification tasks, therefore, conventional method papers use supervised machine learning (CNN, SVM) algorithms for semantic concept or scene detection by training them with extracted feature descriptors [1] [3] [4] [9].

Automated semantic scene or concept detection of a large video is an essential process for video indexing, retrieval and content-based video browsing tasks and previous works have shown efficiency in detecting this. However, not only automated semantic concept or scene detection is important, but how many semantically relevant scenes a large video contains, that is also important for making video indexing and retrieval tasks more efficient and effective which none of the conventional method papers attempted to do to the best of my knowledge.

This paper proposes a new approach to find out semantically relevance between scenes from a large video. This paper aims to detect shot boundaries from a large video, thereafter

selecting candidate key frame, followed by semantic feature extraction known as descriptors, and finally detecting semantic relevance between scenes using deep learning methods from the video dataset, described in section II. Section III explains experimental results of the proposed method and finally section IV concludes the effectiveness of the proposed method.

II. METHODOLOGY

In this section, proposed methodology is explained where CNN-deeper LSTM based model is proposed. CNN is used for feature extraction, thereafter two layers of LSTM are used for scene position detection and semantic relevance detections between scenes. Two layers of LSTM (long short-term memory) are used, thus called deeper LSTM. The system architecture is divided into four steps [Fig. 1] as follows.

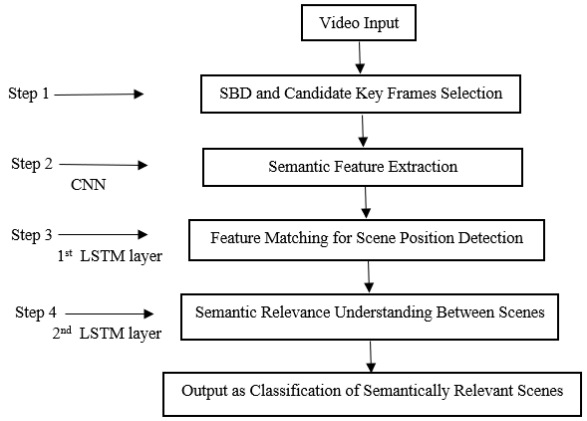


Fig. 1 System Architecture of proposed method

A. SBD and Candidate Key Frame Selection

This is the first step of proposed method which explains candidate key frame selection process. Firstly, shot boundaries are detected from the video in order to segment the video into shots. The method for shot boundary detection (SBD) is used from our previous paper [10]. The summary of our SBD method [10] is using PCA (principal component analysis) for frame feature extraction, followed by using an inter-frame distance (dissimilarity) calculating algorithm to detect shot boundaries. In this paper, video input dataset consisting Cut and Gradual (dissolve) shot boundaries are only used. After shot boundaries are detected, now we have shots of the video, and only the starting and ending frames of each shot are selected as candidate key frames [Fig. 2].

The reason why only starting and ending frames of each shot are selected is that, the ending frames of one shot and the starting frames of the adjacent shot have different frame features, therefore feature dissimilarity can be detected which is described in next section.

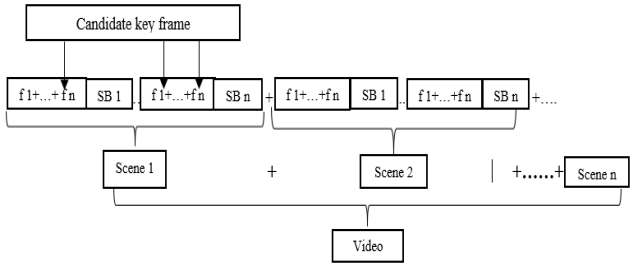


Fig. 2 Candidate Key Frame Selection ('SB' denoted shot boundary, 'f' denotes frame numbers)

B. Semantic Feature Extraction

This section explains about semantic feature extraction. Traditional methods for feature extraction use low-level visual feature extraction, resulting in semantic gap between low-level and high-level semantic meanings. To overcome this gap, deep learning algorithm can be used for semantic feature extraction. Therefore, CNN (convolutional neural network) is used for semantic feature extraction from the candidate key frames of each shot. CNN is well-known for a promising deep learning algorithms for multiple feature extraction. ResNet50 convolutional neural network is used in this paper. These extracted semantic features are called feature descriptors, which are used for scene detection described in next step.

C. Feature Matching for Scene Position Detection

After storing the feature descriptors, next issue arises about the position of the scenes detection. It means, the transformation from one to another scene, in order to find out that, the feature dissimilarity is calculated by using feature matching technique. First layer of the LSTM is trained and then feature descriptors are used for testing. This process classifies feature descriptors in two classes, one is similar key frames and the other one is dissimilar key frames. All the adjacent similar frames are considered as one scene, and after one scene is completed, the system detects a dissimilarity which is considered as next scene, and so on. For example, feature descriptors of one scene have similar feature matching, however, when the next scene begins, there is a dissimilarity between the ending frame feature descriptors of previous scene and the starting frame feature descriptors of the next scene. After detecting scene position, second layer of LSTM is used for semantic relevance understanding between scenes described in next step.

D. Semantic Relevance Understanding Between Scenes

In this step, semantic concepts of the scenes are detection and thereafter, correlation between multiple scenes are analyzed. Firstly, second layer of LSTM is trained with ground truth data (details in section III), and then only the middle frame feature descriptors of each scene are tested for semantically relevant concept detection. Second layer LSTM classifies all the scenes into relevant scenes and irrelevant

scenes. Our video dataset contains multiple scenes of similar concepts, for example, one video sequence contains field sport scenes, water sport scenes and other scenes alternatively. Second layer LSTM detects different sport scenes concept, and then classifies them with their scene positions (e.g., scene 1, scene 2, and so on), hence total number of scenes from each category can be evaluated.

The effectiveness of the proposed method is described in the experimental result section III.

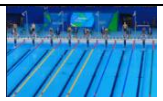
III. Experimental Result

A. Dataset

Our dataset contains total 91 videos, in which 65 videos are used as training dataset and 26 videos are used as testing dataset. The dataset is about Olympic sports collected from online sources. The time span of each video from the dataset is 2-3 minute, and each video consists of multiple similar sport categories. 26 videos contain total 187 scenes. All experiments are conducted on Nvidia GTX 1650 GPU with Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz, running a Windows 10 and MATLAB 2021.

The ground truth data contains four categories. Some examples of the video frames from each category are shown in Table I.

Table I Example of Ground Truth Data

Field sports	Water sports	Ground sports	Court sports
			
			

B. Feature Extraction from Candidate Key Frames

The video dataset contains cut and gradual dissolve shot boundary transition and the SBD method from our previous paper [10] is well suited for detecting these two types of shot boundaries.

This SBD method has given overall 92% accuracy after testing with the Olympic sports video dataset. Beginning and ending frames of each shot are considered as candidate key frame and features are extracted from these fames which reduces time consumption.

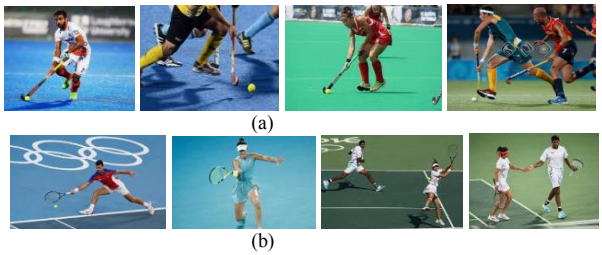
ResNet50 CNN is used for feature extraction. The features from the convolutional layer are stored as feature descriptors.

C. Semantically relevant scene detection

The features descriptors are used for feature matching by using LSTM. LSTM is an algorithm of recurrence neural network (RNN) which can process single point data such as images along with sequences which makes it a perfect algorithm for video processing tasks.

Deeper LSTM or two layers of LSTM is used for scene position detection (first layer) and therefore semantic relevance concept understanding between scenes (second layer). LSTM layers are trained with ground truth dataset shown in Table I and then video dataset is classified into four categories of semantically relevant scenes. The detection results of video number 1, 15 and 26 are given in Table II. Table III represents the detection results of total number of scenes in the video dataset.

The false detection occurred because of not having much feature differences between different scene concepts, shown in Fig. 3. These scenes are having different concepts; however, their background features are similar, also colors are similar as well, thus leading to false detection. Hence, an efficient candidate key frame selection can resolve this issue up to some extent. Our candidate key frame selection strategy selects multiple frames from a single shot as feature descriptors which enables our system to reduce the false alarm rate due to similar background features.



(a) (b)

Fig. 3 Frames having plenty of similar features (a) Field sports, (b) Court sports

Table II Semantically Relevant Scene Detection Result

Video no.	Categories	No. of scenes in actual	No. of scenes detected
1	Field sports	11	6
	Water sports	5	4
	Court sports	9	7
	Ground sports	12	8
15	Field sports	8	6
	Water sports	4	3
	Court sports	6	4
	Ground sports	10	7
26	Water sport	6	5
	Court sports	8	5
	Ground sports	14	7

The field sport and ground sports accuracy are comparatively lower than the other two categories, it's because of having similar features in presence of fast motion activity.

Table III Detection Accuracy

Total Scenes	Field sports	Water sports	Court sports	Ground sports
Actual	64	26	53	44
Detected	52	23	47	35
Detection accuracy (%)	81.25%	88.4%	88.6%	79.54%

The overall experimental results have significantly achieved good accuracy in detecting the semantically relevant scenes from the video dataset.

IV. CONCLUSIONS

In conclusion, this paper has proposed a new approach for semantically relevant scene detection based on CNN-deeper LSTM model by detecting the inter correlation between multiple scenes in a video. Even though, this method has found some false detection between the categories with similar features, the overall accuracy is well achieved which shows the effectiveness of the proposed method. Detection of concepts of the scene and thereby categorizing scenes consisting relevant scenes from a video increase the efficiency and reduce time consumption in video indexing, retrieval and content-based video searching tasks.

For future work, semantically relevant complex scenes of a video can be detected, i.e., scenes which have plenty of similar features and thus difficult to recognize, such as running and throwing javelin, both sports have similar ground background, so detection of these scenes by categorizing in separate groups can be improved. Another future work could be recognition of appearance and activity of a specific person recognition from a large video.

ACKNOWLEDGMENT

The research study has been financially supported by Petchra Pra Jom Klao Research Scholarship from King Mongkut's University of Technology Thonburi. I want to thank my mother Mrs. Jhulan Chakraborty and my father Mr. Prabir Chakraborty for their continuous support and valuable advices to me. I also want to thank my laboratory-mate for helping in finding solution of my certain research problems.

REFERENCES

[1] F. Markatopoulou, V. Mezaris, N. Pittars, and I. Patras, "Local Features and a Two-Layer Stacking Architecture for Semantic Concept Detection in Video," *in IEEE Transaction on Emerging Topics in Computing*, vol. 3, no. 2, pp. 193-204, June 2015, doi: 10,1109/TETC.2015.241874.

[2] C. Yu, W. Feng, and H. Zhou, "A semantic analysis method based on concept lattice rules streamlined by negative samples," *2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet)*, 2012, pp. 1583-1586, doi: 10.1109/CECNet.2012.6201867.

[3] H. Song, X. Wu, W. Yu, and Y. Jia, "Extracting Key Segments of Videos for Event Detection by Learning from Web Sources," *in IEEE Transaction on Multimedia*, vol. 20, no. 5, pp. 1088-1100, May 2018, doi: 10,1109/TMM.2017.2763322.

[4] H. Min, W. De Neve, and Y. M, "Towards Using Semantic Features for Near-duplicate Video Detection," 2010 *IEEE International Conference on Multimedia and Expo*, pp. 1364-1369, 2010, doi: 10,1109/ICME.2010.5583219.

[5] J. Luo, C. Papin, and K. Costello, "Towards Extracting Semantically Meaningful Key Frames from Personal Video Clips: from Human to Computers," *in IEEE Transaction on Circuits and System for Video Technology*, vol. 19, no. 2, pp. 289-301, Feb. 2009, doi: 10,1109/TCSVT.2008.2009241.

[6] A. Savakis and A. M. Shringarpure, "Semantic Background Estimation in Video Sequences," *2018 IEEE 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 597-601, Feb. 2018, doi: 10,1109/SPIN.2018.8474279.

[7] Han Xiuli, Wu Lifang, Liu Xingsheng, Cheng Zhaohui and Gong Yu, "Offense-defense semantic analysis of basketball game based on motion vector," *2009 International Conference on Image Analysis and Signal Processing*, 2009, pp. 146-149, doi: 10.1109/IASP.2009.5054655.

[8] Gu Xu, Yu-Fei Ma, Hong-Jiang Zhang and Shi-Qiang Yang, "An HMM-based framework for video semantic analysis," *in IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 11, pp. 1422-1433, Nov. 2005, doi: 10.1109/TCSVT.2005.856903.

[9] M. Soltanian and S. Ghaemmaghami, "Hierarchical Concept Score Postprocessing and concept-Wise Normalization in CNN-Based Video Event Recognition," *in IEEE Transaction on Multimedia*, vol. 21, no. 1, pp. 157-172, Jan. 2019, doi: 10,1109/TMM.2018.2844101.

[10] D. Chakraborty, W. Chiracharit and K. Chamnongthai, "Video Shot Boundary Detection Using Principal Component Analysis (PCA) and Deep Learning," *2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 2021, pp. 272-275, doi: 10.1109/ECTI-CON51831.2021.9454775.